4-2016

Multilevel IRT: When is local independence violated?

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Jessica Jacovidis
James Madison University

Multilevel Item Response Theory (IRT): When is Local Independence Violated?

Christine E. DeMars and Jessica Jacovidis

James Madison University

(April, 2016). Poster presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Multilevel Item Response Theory (IRT): When is Local Independence Violated?

One assumption for item response theory (IRT) parameter estimation is local independence: Conditional on the item and person parameters, item responses should be independent. Measurement students learn that the clustering of abilities (or other outcome variables) within schools violates independence of errors in regression models; these students may wonder whether clustering effects violate local independence in IRT. The purpose of this study is to explain and illustrate that conditioning on ability eliminates the within-school correlation of responses if the within-school correlation is due to similar abilities. For this reason, testing companies generally do not model students nested within schools when calibrating items. A secondary purpose is to acknowledge and illustrate that clustering of students within schools may violate local independence if the schools introduce random item effects. No argument is made here about whether these school-related item effects occur in practice; the purpose is instead to help develop an understanding of local independence and to concede that clustering might lead to violations, albeit not due to the clustering of abilities within schools.

For research on the relationships between test scores and other outcomes, clustering of student scores (abilities) within schools must be taken into account when estimating the structural part of the model (latent regression[1]), or regression coefficients/mean group differences for models without latent variables. For example, Hedges and Hedberg (2007) studied several standardized math and reading tests for grades K-12, and reported intraclass correlation coefficients (ICCs) mostly in the range of 0.2 to 0.3. If ignored, this clustering can lead to

[1] For examples of latent regression in a multilevel framework, see Kamata (2001) or Pastor (2003).

underestimated standard errors of the regression coefficients. In contrast, in the measurement part of the model, nesting of students within schools may be subsumed within the person parameter, θ (ability). Even if a sizable portion of the variance in θ is at the school level, the responses could be locally independent, conditional on θ. Thus, the nesting of students within schools is often ignored when calibrating statewide testing data. This is a mathematical tautology: conditioning on ability at the student level controls for school differences in ability. This study will provide a concrete illustration of this principle.

It should be acknowledged, however, that it is possible for responses to an item to be more similar within schools than between schools, even after conditioning on θ. This is analogous to nesting of items within schools; there is a random school effect for item difficulty.[2] If such an effect exists and is not modeled, the local independence assumption will be violated and the standard errors of the item parameters will be misestimated. This random effect is a school by item interaction, so it is not removed by conditioning on ability and item difficulty. In this paper, we will not attempt to investigate how common this problem is in practice.[3] Rather, an illustration of the potential for local dependence will be provided, as contrast to the more common situation where the focus is on abilities nested within school. This example is not intended to systematically explore the effects of each factor; rather it is intended to concretely illustrate a theoretical point.

[2] This is unrelated to the random-effect models due to clustering of items within testlets or other bundles (such as the random effects models in Scott & Ip, 2002; or Wainer, Bradlow, Wang, 2007). Clustering of items within testlets is a classic violation of local independence and is unrelated to clustering within schools.

[3] Cohen, Chan, Jiang, and Seburn (2008) suggested "Every real data set exhibits substantial intraclass correlation in specific item responses" (p. 309). Here, we are simply allowing the possibility of this ICC, not investigating its prevalence.

Method

Data were simulated under three conditions: 1) no random school effects); 2) random school effect for θ with intraclass correlation coefficient (ICC) = .30; and 3) random school effect for item difficulty with ICC = .04.[4]

The model was: $P_i(\theta_{jk}) = c_i + (1-c_i)\frac{e^z}{1+e^z}$, where $z = a_i[(\theta_{jk} - b_i]$ for condition 1, $z = a_i[r_{jk} + u_k - b_i]$ for condition 2, or $z = a_i[\theta_{jk} - (b_i + v_{ik})]$ for condition 3. Subscript *i* indicates item *i*, subscript *j* indicates person *j*, and subscript *k* indicates school *k*. a_i is the item discrimination, b_i is the item difficulty, c_i is the lower asymptote. In conditions 1 and 3, θ_{jk} is the ability of student *j* within school *k*. In condition 2, r_{jk} is the random effect for student *j* within school *k* and u_k is the random effect for school ability; $r_{jk} + u_k$ is = θ_{jk}. In condition 3, v_{ik} is the random effect of school *k* on item *i*'s difficulty. Looking at these equations, it should be obvious why there is no need to model student abilities as nested within schools (unless modeling additional coefficients in a latent regression model): θ_{jk} subsumes the random school effect u_k, so the response residuals are uncorrelated within schools. The example below simply illustrates this theoretical point empirically.

In each of 1000 replications, 20 schools were simulated, with 100 students in each school. θs were ~N(0,1); in condition 2 this meant that the within-school standard deviation = $\sqrt{7}$ and the between-school standard deviation = $\sqrt{3}$. The test had 40 items. The data followed a 3PL model, with *c* = .15, *a* = 1.7, and *b* ranging from -1.51 to 1.51, with spacing to reflect a normal distribution. For condition 3, a random school effect was added to each *b*, ~N(0,.02), to produce the desired

[4] Phillips (2015) found an average ICC = .04 for item difficulties in one testing program.

ICC.[5] Flexmirt 2.01 (Cai, 2013) was used to calibrate the item parameters through MML, followed by expected-a-posterior (EAP) estimates of the person parameters. Priors were set on the logit-*g* ($N(-2.197,0.5^2)$) and the *a* ($\log N(0.3,0.5^2)$).

Results

The *b*-parameters had a small negative bias except for the most difficult items. The *c*-parameters were nearly unbiased, although the bias became less negative as item difficulty increased because the prior had less influence as difficulty increased. Notably, any bias in the *b* and *c* parameters was virtually the same regardless of school effects.

Figure 1 shows the bias for the *a*-parameters. The absolute value of the bias was slightly but consistently more negative (median = 9% more) when there was a school effect on the difficulty. Because of this bias, the *d*-parameters (not shown) were biased negatively for easy items and positively for hard items when there was a school effect for item difficulty.

Empirical SEs for the *b*-parameters are shown in Figure 2. The school effect for θ did not impact the SE, but the school effect for item difficulty increased the SE by about 5-20% (median = 14%). SEs for *a*-parameters (Figure 3) and *c*-parameters (Figure 4), however, were not greatly influenced by school effects.

Bias and empirical SE for the θs are shown in Figures 6 and 7. The absolute value of the bias was slightly larger when there was a random school effect for items, perhaps because the estimated information was lower due to the negatively biased *a*-parameters, which gave more weight to the prior for the θs. The SE was considerably larger. Using the standard errors in the

[5] Random school effects were not correlated across items. Correlated random effects would produce classic violation of local independence.

output, the reliability estimate was .92 for the conditions of no school effect and school effect for θ, which was slightly larger than $\hat{\rho}^2_{\theta\hat{\theta}}$ = .90. However, the standard errors in the output were too low when there was a school effect for item difficulty, producing an estimate of .92 when $\hat{\rho}^2_{\theta\hat{\theta}}$ = .63.

Summary and Implications

The practical implication of these findings is that the well-established issue of nesting of student abilities within schools is not a problem for the SEs in the measurement part of the model (IRT calibration). A fairly large school effect for ability (ICC = .30) did not bias the parameters or SEs beyond what was seen in the data with no clustering. However, school effects on item difficulty, a less-studied (and hopefully less prevalent) problem, increased the SEs of both item difficulty and ability parameter estimates. This study demonstrated this issue with a small ICC of .04. School effects on the item difficulties may be viewed as a violation of parameter invariance. The item parameters depend on the school, and if the same parameters are used for all schools, the responses violate local independence, biasing the estimated item discriminations and inflating the standard errors of the other parameters. In some sense, this is differential item functioning (DIF), but unless the random effects are correlated with school characteristics, it is unlikely to be detectable in DIF analyses.

The impact of the random school effect for item difficulty should decrease as the number of schools increases, and thus might not be an issue if the item parameters are calibrated with the complete operation data. However, pilot test items and linking items are often distributed across several forms of the test. For security purposes, only two or three forms may be spiraled within

each school, which means that each form is completed by a large number of students in a relatively small number of schools. Additionally, members of some small demographic groups may be clustered in a few schools in one region of a state. Thus, although clustering of student abilities within schools does not violate local independence, the potential for violations due to random school effects on items cannot be immediately dismissed.

If researchers want to go beyond the IRT calibration and estimate group differences or regression parameters, multilevel (random effects) models are needed to get correct estimates of the standard errors. Typically that is not part of the item calibration, but the broader picture should not be overlooked.

References

Cai, L. (2013). flexMIRT version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring. [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cohen, J., Chan, T., Jiang, T., & Seburn, M. (2008). Consistent estimation of Rasch item parameters and their standard errors under complex sample designs. *Applied Psychological Measurement, 32*, 289-310.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized experiments in education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79-93.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16,* 223-243.

Phillips, G. W. (2015). Impact of design effects in large-scale district and state assessments. *Applied Measurement in Education, 28*, 33-47.

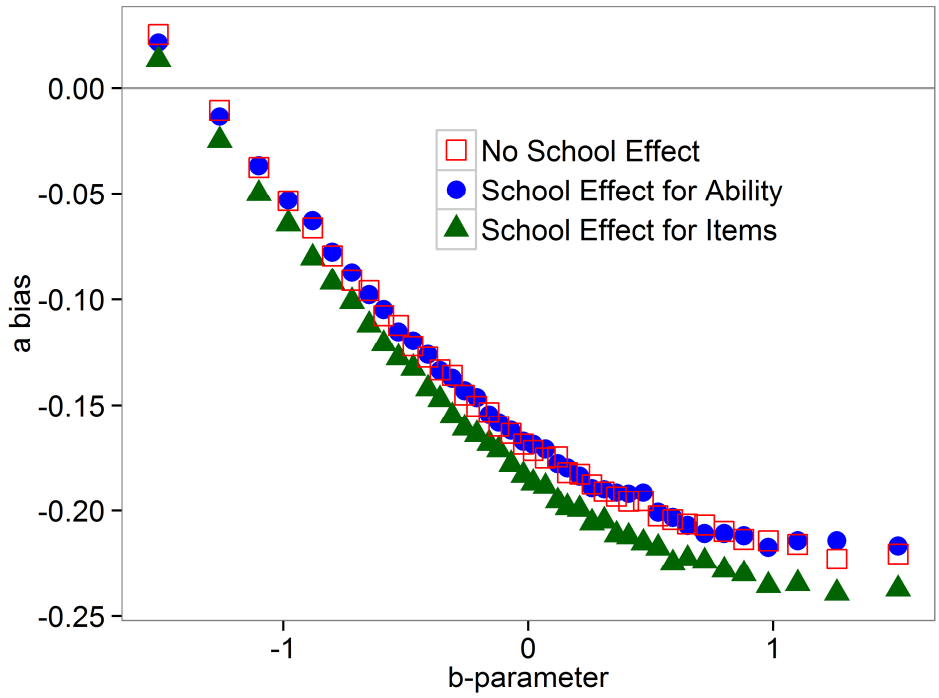

Figure 1: Bias in *a*-parameters.

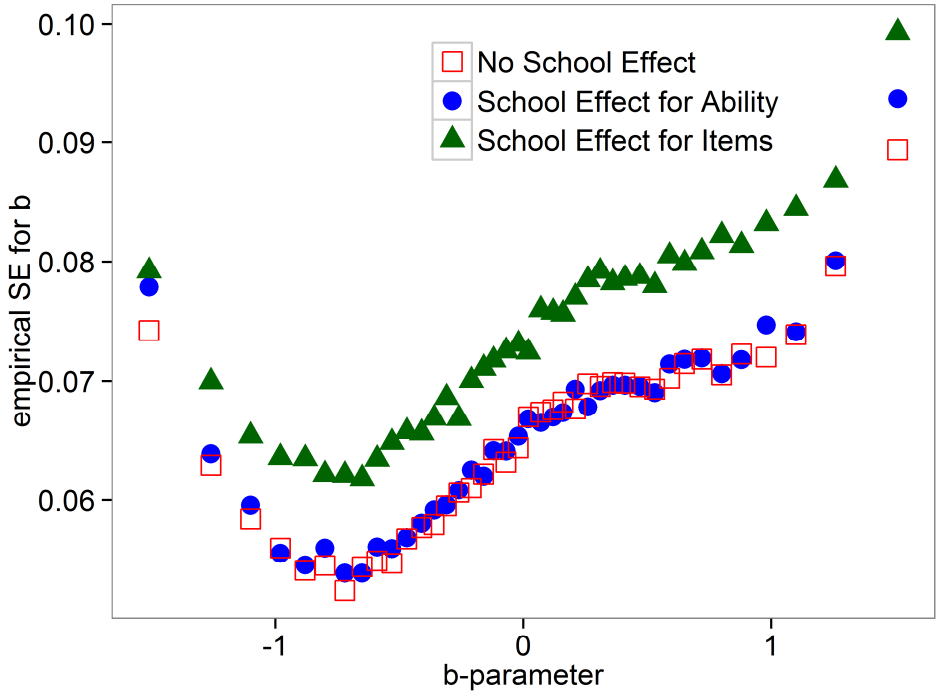

Figure 2: Empirical SE of *b*-parameters.

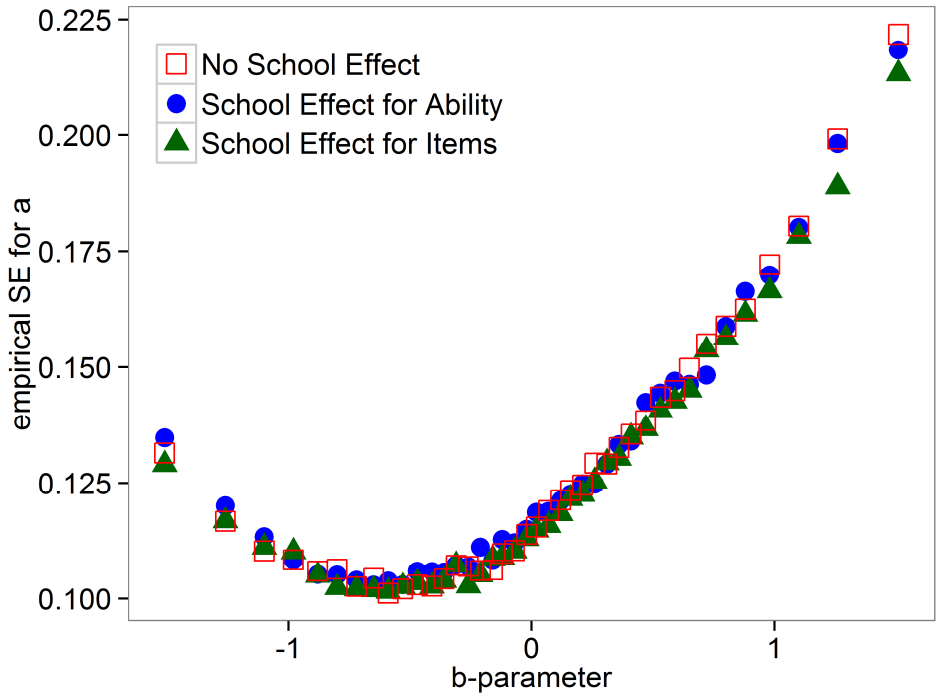

Figure 3: Empirical SE of *a*-parameters.

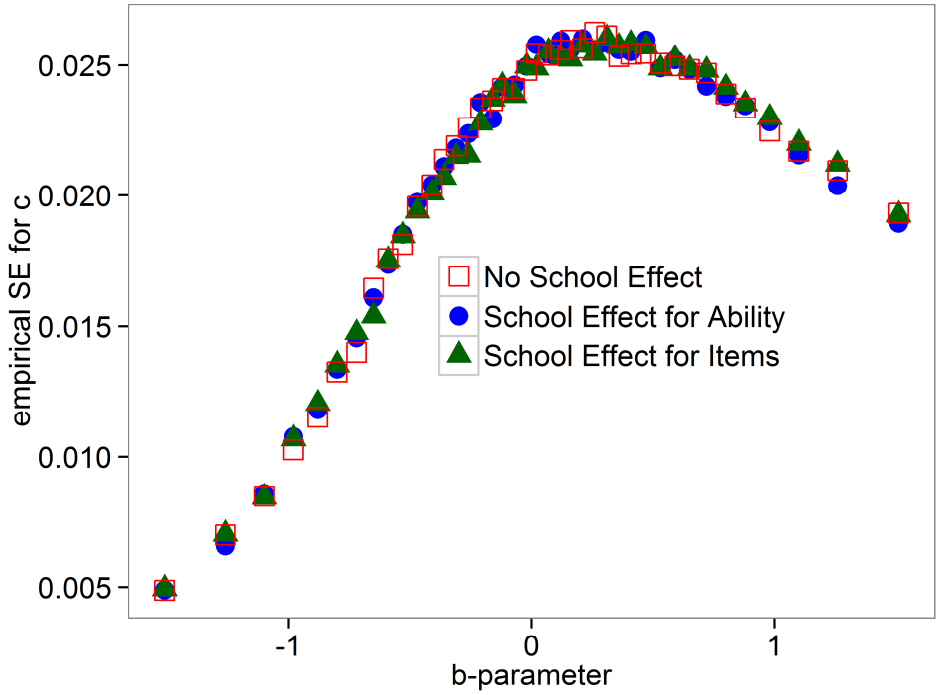

Figure 4: Empirical SE of *c*-parameters.

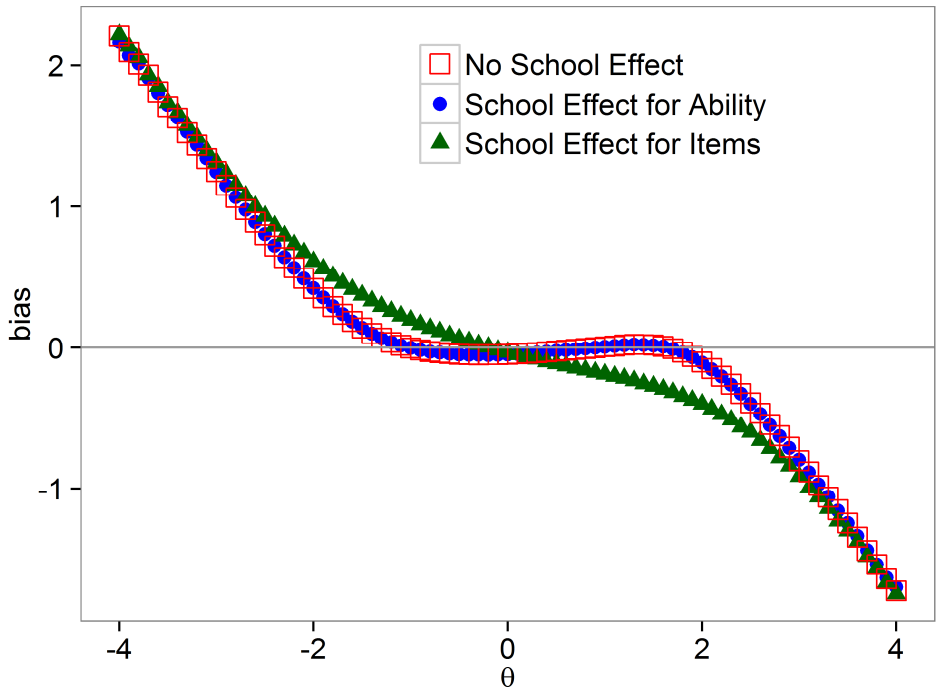

Figure5: Bias in θ-parameters.

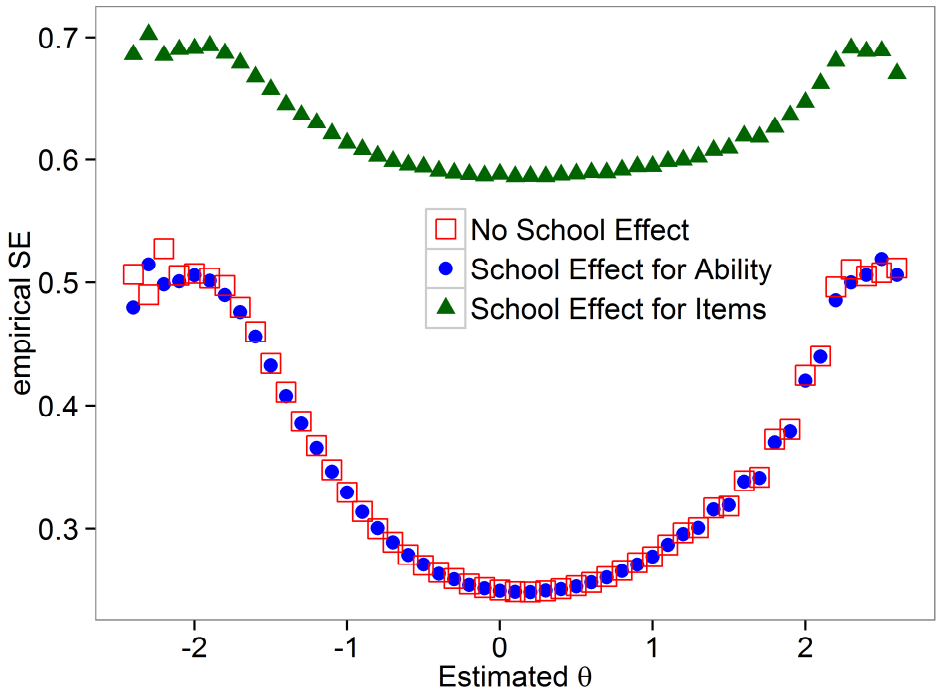

Figure 6: Empirical SE of θ-parameters.